Big Data Analytics for Manufacturing Internet of Things: Opportunities, Challenges and Enabling Technologies

Hong-Ning Dai, Hao Wang, Guangquan Xu, Jiafu Wan, Muhammad Imran

ARTICLE HISTORY
Compiled June 16, 2019

ABSTRACT
The recent advances in information and communication technology (ICT) have promoted the evolution of conventional computer-aided manufacturing industry to smart data-driven manufacturing. Data analytics in massive manufacturing data can extract huge business values while can also result in research challenges due to the heterogeneous data types, enormous volume and real-time velocity of manufacturing data. This paper provides an overview on big data analytics in manufacturing Internet of Things (MIoT). This paper first starts with a discussion on necessities and challenges of big data analytics in manufacturing data of MIoT. Then, the enabling technologies of big data analytics of manufacturing data are surveyed and discussed. Moreover, this paper also outlines the future directions in this promising area.

KEYWORDS
Smart Manufacturing; Data Analytics; Data Mining; Internet of Things

H.-N. Dai is with Faculty of Information Technology, Macau University of Science and Technology, Macau. E-mail: hndai@ieee.org.

H. Wang is with Faculty of Engineering and Natural Sciences, Norwegian University of Science & Technology, Gjøvik, Norway. Email: hawa@ntnu.no

G. Xu is with Tianjin Key Laboratory of Advanced Networking,), College of Intelligence and Computing, Tianjin University, Tianjin, China Email: losin@tju.edu.cn

J. Wan is with School of Mechanical & Automotive Engineering, South China University of Technology, Guangzhou, China Email: mejwan@scut.edu.cn

M. Imran is College of Computer and Information Sciences, King Saud University, Saudi Arabia, Email: dr.m.imran@ieee.org

1. Introduction

The manufacturing industry is experiencing a paradigm shift from automated manufacturing industry to "smart manufacturing" [Kusiak(2018)]. During this evolution, Internet of Things (IoT) plays an important role of connecting the physical environment of manufacturing to the cyberspace of computing platforms and decision-making algorithms, consequently forming a Cyber-Physical System (CPS). We name such industrial IoT dedicated to manufacturing industry as manufacturing IoT (MIoT) in this paper.

MIoT consists of a wide diversity of manufacturing equipments, sensors, actuators, controllers, RFID tags and smart meters, which are connected with computing platforms through wired or wireless communication links. There is a surge of big volume of data traffic generated from MIoT. The MIoT data is featured with large volume, heterogeneous types (i.e., structured, semi-structured, unstructured) and is generated in a real-time fashion. The analytics of MIoT data can bring many benefits, such as improving factory operation and production, reducing machine downtime, improving product quality, enhancing supply chain efficiency and improving customer experience [Zhong et al.(2017), Lade, Ghosh, and Srinivasan(2017), Tao et al.(2018)]. However, there are also many challenges in data analytics in MIoT in the different phases of the whole life cycle of data analytics.

There are several surveys on data analytics in manufacturing industry. The work of [Tao et al.(2018)] proposes a data-driven smart manufacturing framework and provides several application scenarios based on this conceptual framework. The necessities of big data analytics in smart manufacturing are summaried in [Kusiak(2017)]. The work of [Lade, Ghosh, and Srinivasan(2017)] provides an overview on data analytics in manufacturing with a case study. Tao and Qi presents an overview of service-oriented manufacturing in [Tao and Qi(2019)]. However, most of the aforementioned studies lack of the introduction of enabling technologies corresponding to the challenges, which are of interest to both academic researchers and industrial practitioners.

Therefore, the aim of this paper is to provide an overview on data analytics in MIoT from opportunities, challenges and enabling technologies. The main contributions of this paper can be summarized as follows.

- We provide a summary on key characteristics of MIoT and a life cycle of big data analytics for MIoT data. We also discuss necessities and challenges of big data analytics in MIoT.
- We present an overview on enabling technologies of big data analytics for MIoT from the aspects of data acquisition, data preprocessing and data analytics.
- We given an outline of future research directions in aspects of security, privacy, fog computing and new data analytics methods.

The rest of this paper is organized as follows. Section 2 gives the discussion on necessities and challenges of big data analytics in MIoT. Section 3 introduces enabling technologies of big data analytics in MIoT. Section 4 discusses the future research directions. Finally, this paper is concluded in Section 5.

Table 1. Comparison between MIoT and CIoT

	Manufacturing IoT	**Consumer IoT**
Goal	Manufacturing-industry Centric	Consumer Centric
Devices	Machines, Sensors, Controllers, Actuators, Smart meters	Consumer devices and Smart appliances
Working Environment	Harsh (vibration, noisy, extremely high/low temperature)	Moderate
Data rate	High (usually)	Low or average
Delay	Delay sensitive	Delay tolerant
Mission	Mission-critical	Non-mission-critical

2. Necessities and challenges of big data analytics for Manufacturing Internet of Things

In this section, we first introduce the key characteristics of Manufacturing Internet of Things in Section 2.1. We then introduce the life cycle of big data analytics for MIoT in Section 2.2. We next discuss the necessities of big data analytics for MIoT in Section 2.3 and the challenges in Section 2.4.

2.1. *Key characteristics of Manufacturing Internet of Things*

In this paper, we roughly categorize IoT into consumer Internet of Things (CIoT) and Manufacturing Internet of Things (MIoT). Table 1 compares MIoT with CIoT. In contrast to MIoT, CIoT mainly serve for consumers. Hence, CIoT mainly consists of consumer devices (e.g., smart phones, wearable electronics) and smart appliances (e.g., refrigerators, TVs, washing machines). CIoT mainly aims to improve user experience while MIoT mainly focuses on improving factory operations and production, reducing the machine downtime and improving product quality. Moreover, MIoT usually works in harsh industrial environment (like vibrated, noisy and extremely high/low temperature) while CIoT works in moderate environment. In addition, MIoT applications usually require high data-rate network connection with low delay while CIoT applications have relaxed requirement on network connection. Furthermore, MIoT systems are usually mission-critical and sensitive to system failure or machinery downtime while CIoT systems are non-mission-critical.

In this paper, we mainly focus on MIoT. The MIoT ensures the connection of various *things* (smart objects) mounted with various electronic or mechanic sensors, actuators, instruments and software systems which can sense and collect information from the physical environment and then make actions on the physical environment. During this procedure, the data analytics plays an important role in extracting informative values, forecasting the coming events and predicting the increment/decrements of products.

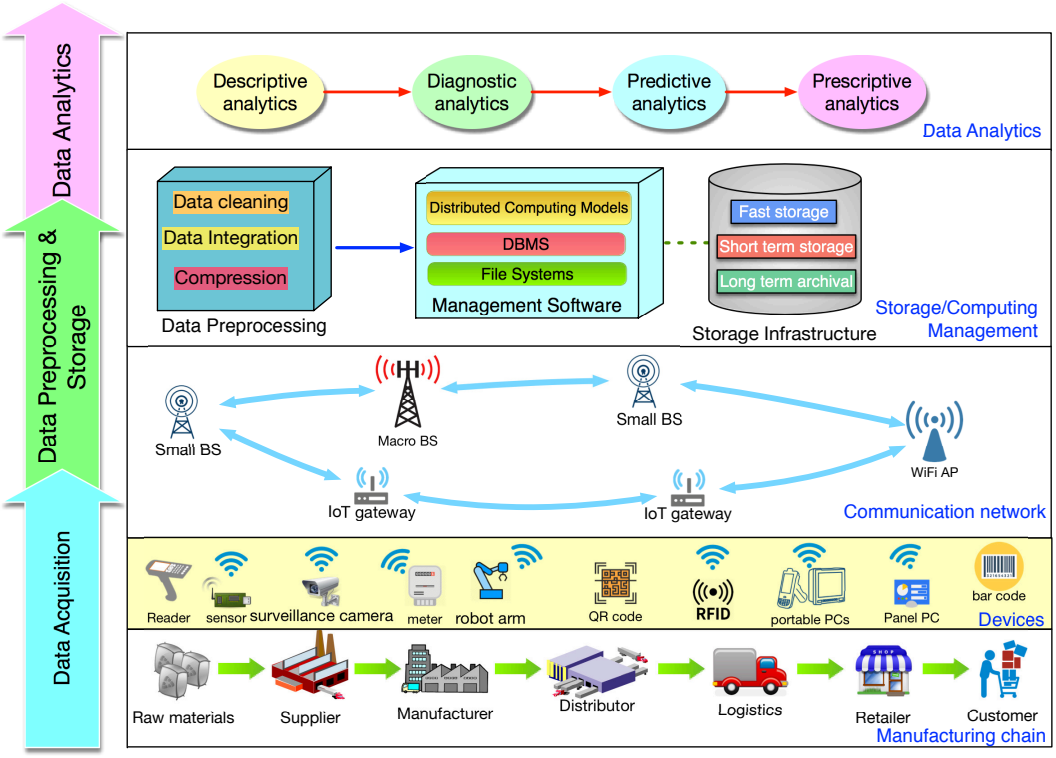


Figure 1. Life cycle of Big Data Analytics for MIoT

2.2. *Life cycle of big data analytics for MIoT*

We first introduce the life cycle of big data analytics for MIoT. Figure 1 shows that the life cycle of big data analytics for MIoT consists of three consecutive stages: 1) Data Acquisition, 2) Data Preprocessing and Storage, 3) Data Analytics. There are other taxonomies [Hu et al.(2014), Casado and Younas(2015), Tao et al.(2018)]. We categorize the life cycle of big data analytics into the above three stages since this taxonomy can accurately capture the key features of big data analytics in MIoT.

1. *Data acquisition* consists of data collection and data transmission. Firstly, data collection involves acquiring raw data from various data sources in the whole manufacturing process via dedicated data collection technologies. For example, RFID tags are scanned by RFID readers in product warehouse. Then, the collected data will be transmitted to the data storage system through either wired or wireless communication systems. Details about enabling technologies of data acquisition are given in Section 3.1.
2. *Data preprocessing and storage.* After data collection, the raw data needs to be preprocessed before keeping them in data storage systems because of the big volume, redundancy, uncertainty features of the raw data [Lade, Ghosh, and Srinivasan(2017)]. The typical data preprocessing techniques include data cleaning, data integration and data compression. Data storage refers to the process of storing and managing massive data sets. We divide the data storage system into two components: storage infrastructure and data management software. The infrastructure not only includes the storage devices but also the network devices connecting the storage devices together. In addition to the networked storage

devices, data management software is also necessary to the data storage system. Details about enabling technologies of data preprocessing and data storage are given in Section 3.2.

3. *Data analytics.* In data analysis phase, various data analytical schemes are used to extract valuable information from the massive manufacturing data sets. We roughly categorize the data analytical schemes into four types: (i) statistic modelling, (ii) data visualization, (iii) data mining and (iv) machine learning. Details about enabling technologies of data analysis are presented in Section 3.3.

2.3. *Necessities of big data analytics for MIoT*

There is an enormous amount of data generated from the whole manufacturing chain consisting of raw material supply, manufacturing, product distribution, logistics and customer support, as shown in Figure 1. Such "big data" needs to be extensively analysed so that some valuable and informative information can be extracted.

We summarize the reasons of big data analytics for MIoT as follows:

- *Improving factory operations and production.* The predictive analytics of manufacturing data and customer demand data can help to improve machinery utilization consequently enhancing factory operations. For example, the demands for certain products are often related to weather or seasonal conditions (e.g., down coats related to the cold weather). Forecasting a cold wave can be used to make pro-active allocation of machinery resources and pre-purchasing raw materials to fulfill the upsurge demands.
- *Reducing machine downtime.* The prevalent sensors deployed throughout the whole product line can collect various data reflecting machinery status. For example, the analysis of machinery health data can help to identify the root cause of failure consequently reducing machine downtime [Lade, Ghosh, and Srinivasan(2017)]. Moreover, the sensory data from automatic assembly line can also be used to determine excessive load of machines so as to balance the loads among multiple machines [Wang et al.(2018a)].
- *Improving product quality.* On one hand, the analysis of market demand and customer requirement can be used to improve the product design in reflecting product improvements. During the product manufacturing procedure, the analysis of manufacturing data can help to reduce the ratio of defective goods by identifying the root cause. As a result, the product quality can be improved.
- *Enhancing supply chain efficiency.* The proliferation of various sensors, RFID and tags during supplier, manufacturing and transportation generates massive supply chain data, which can be used to analyse supply risk, predict delivery time, plan optimal logistic route, etc. Moreover, the analysis of inventory data can reduce the holding costs and fulfill the dynamic demands by establishing safety stock levels. In addition, big data analytics on IoT-enabled intelligent manufacturing shops [Zhong et al.(2017)] can also help to make accurate logistic plan and schedules. As a result, the system efficiency can be greatly improved.
- *Improving customer experience.* Companies can obtain customer data from various sources, such as sales channels, partner distributors, retailers, social media platforms. Then, big data analytics on customer data offers descriptive, predictive and prescriptive solutions to enable companies to improve product design, quality, delivery, warrant and after-sales support. As a result, the customer experience can be improved. For example, the IoT data in the whole food supply-chain

is also beneficial to prevent mischievous actions and guarantee food safety [Leng et al.(2018b)].

2.4. *Challenges of big data analytics for MIoT*

MIoT data has the following characteristics: (1) massive volume, (2) heterogeneous data types, (3) being generated in real-time fashion and (4) bringing huge both business value and social value. The unique features cause the research challenges in big data analytics for MIoT. We summarize the challenges in the following aspects.

1. Challenges in data acquisition

Data acquisition addresses the issues including data collection and data transmission, during which there are the following challenges.

- *Difficulty in data representation.* MIoT data has different types, heterogeneous structures and various dimensions. For example, manufacturing data can be categorized into structured data, semi-structured and un-structured data [Tao et al.(2018)]. How to represent these structured, semi-structured and unstructured data becomes one of major challenges in big data analytics for MIoT.
- *Efficient data transmission.* How to transmit the tremendous volumes of data to data storage infrastructure in an efficient way becomes a challenge due to the following reasons: (i) *high bandwidth consumption* since the transmission of big data becomes a major bottleneck of wireless communication systems [Hu et al.(2014)]; (ii) *energy efficiency* is one of major constraints in many wireless industrial systems, such as industrial wireless sensor networks [Azoidou et al.(2017)].

2. Challenges in data preprocessing and storage

Data generated from MIoT leads to the following research challenges in data preprocessing.

- *Data integration.* Data generated in MIoT has the various types and heterogeneous features. It is necessary to integrate the various types of data so that efficient data analytics schemes can be implemented. However, it is quite challenging to integrate different types of MIoT data.
- *Redundancy reduction.* The raw data generated from MIoT is characterized by the temporal and spatial redundancy, which often results in the data inconsistency consequently affecting the subsequent data analysis. How to mitigate the data redundancy in MIoT data becomes a challenge.
- *Data cleaning and data compression.* In addition to data redundancy, MIoT data is often erroneous and noisy due to the defected machinery or errors of sensors. However, the large volume of the data makes the process of data cleaning more challenging. Therefore, it is necessary to design effective schemes to compress MIoT data and clean the errors of MIoT data.

Data storage plays an important role in data analysis and value extraction. However, designing an efficient and scalable data storage system is challenging in MIoT. We summarize the challenges in data storage as follows.

- *Reliability and persistency of data storage.* Data storage systems must ensure the reliability and the persistency of MIoT data. However, it is challenging to fulfill the above requirements of big data analytics while balancing the cost due to the tremendous amount of data [Guerra et al.(2011)].

- *Scalability.* Besides the storage reliability, another challenging issue lies in the scalability of storage systems for big data analytics. The various data types, the heterogeneous structures and the large volume of massive data sets of MIoT lead to the in-feasibility of conventional databases in big data analytics. As a result, new storage paradigms need to be proposed to support large scale data storage systems for big data analytics.
- *Efficiency.* Another concern with data storage systems is the efficiency. In order to support the vast number of concurrent accesses or queries initiated during the data analytics phase, data storage needs to fulfill the efficiency, the reliability and the scalability requirements together, which is extremely challenging.

3. Challenges in data analytics

It is quite challenging in big data analytics for MIoT due to the tremendous volume, the heterogeneous structures and the high dimension. The major challenges in this phase are summarized as follows.

- *Data temporal and spatial correlation.* Different from conventional data warehouses, MIoT data is usually spatially and temporally correlated. How to manage the data and extract valuable information from the temporally/ spatially-correlated MIoT data becomes a new challenge.
- *Efficient data mining schemes.* The tremendous volume of MIoT data leads to the challenge in designing efficient data mining schemes due to the following reasons: (i) it is not feasible to apply conventional multi-pass data mining schemes due to the huge volume of data, (ii) it is critical to mitigate the data errors and uncertainty due to the erroneous features of MIoT data.
- *Privacy and security.* It is quite challenging to pertain the privacy and ensure the security of data during the analytics process. Though there are a number of conventional privacy-preserving data analytical schemes, they may not be applicable to the MIoT data with the huge volume, heterogeneous structures, and spatio-temporal correlations. Therefore, new privacy-preserving data mining schemes need to be proposed to address the above issues.

3. Enabling Technologies

In this section, we discuss the enabling technologies of big data analytics in MIoT. According to the three phases in the life cycle of big data analytics in MIoT, we roughly categorize these technologies into data acquisition, data preprocessing and storage, data analytics. In particular, we first discuss the data acquisition related technologies in Section 3.1. We then describe the data preprocessing and storage in Section 3.2. In Section 3.3, we discuss the data analytics in MIoT.

3.1. *Data acquisition*

As shown in Figure 1, the whole manufacturing chain involves with multiple parties such as suppliers, manufacturers, distributors, logistics, retailers and customers. As a result, different types of data sources generate from each of these sectors. Take a manufacturing factory an example. Sensors deployed at the production line can collect device data, product data, ambient data (like temperature, humidity, air pressure), electricity consumption, etc. In the product warehouse, RFID or other tags can help

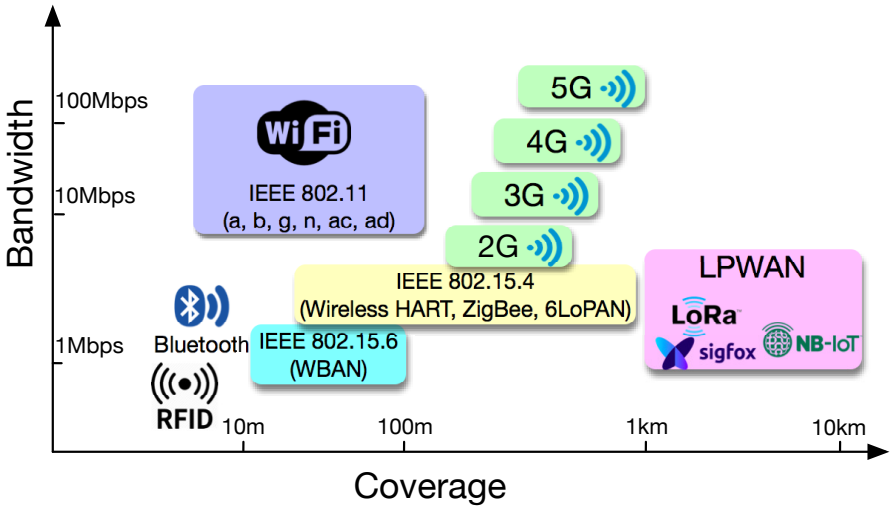


Figure 2. Wireless Communication Technologies for MIoT (figure is not scalable)

to identify and track products. RFID tags attached at products can be read in a short distance by a RFID reader in a wireless manner.

The collected data can then be transmitted to the next stage via either wired or wireless manner. Industrial Ethernet is one of the most typical wired connections in manufacturing. When Ethernet is applied to an industrial setting, more rugged connectors and more durable cables are often required to satisfy harsh environment requirements (like vibration, noise and temperature). Compared with wired communications, wireless communications do not require communication wiring and related infrastructure consequently saving the cost and improving scalability. The major obstacle of the wide deployment of wireless communications in industrial systems is the lower throughput and the higher delay than wired communications. However, the recent advances in wireless communications make wireless connections feasible in industrial components.

Various sensors, RFIDs and other tags can connect with IoT gateways, WiFi Access Points (APs), small base station (BS) and macro BS to form an industrial wireless sensor networks (IWSN) [Chi et al.(2014)]. It is worth mentioning that different wireless technologies have different coverage and bandwidth capabilities. Figure 2 gives the comparison of various wireless technologies regarding to coverage and bandwidth. In particular, it is shown in Figure 2 that conventional wireless technologies like Near Field Communications (NFC), RFID, Bluetooth Low Energy (LE), wireless body sensor networks (WBAN), Internet Protocol (IPv6), Low-power Wireless Personal Area Networks (6LoWPAN) and Wireless Highway Addressable Remote Transducer (WirelessHART) [Petersen and Carlsen(2011)] are suffering from short communication range (i.e., most of them can typically cover less than hundreds of meters). As a result, they cannot support the wide-coverage industrial applications, like smart metering, smart cities and smart grids [Xu et al.(2017)]. It is true that other wireless technologies such as WiFi (IEEE 802.11) and mobile communication technologies (such as 2G, 3G, 4G networks) can provide longer coverage range while they often require high energy consumption at handsets, whereas most of sensor nodes have the limited energy (i.e., supplied by batteries). Therefore, WiFi and other mobile communication technologies may not be feasible in IWSN due to the high energy consumption.

Recently, Low Power Wide Area Networks (LPWAN) essentially provide a solution

to the wide coverage demand while saving energy. Typically LPWAN technologies include Sigfox, LoRa, Narrowband IoT (NB-IoT) [Mekki et al.(2018)]. LPWAN has lower power consumption than WiFi and mobile communication technologies. Take NB-IoT as an example. It is shown in [Xu et al.(2017)] that an NB-IoT node can have a ten-year battery life. Moreover, LPWAN has a longer communication range than RFID, bluetooth and 6LoWPAN. In particular, LPWAN technologies have the communication range from 1km to 10 km. Furthermore, they can also support a large number of concurrent connections (e.g., NB-IoT can support 52,547 connections as shown in [Xu et al.(2017)]). However, one of limitations of LPWAN technologies is the low data rate (e.g., NB-IoT can only support a data rate upto 250 kps). Therefore, LPWAN technologies should complement with conventional RFID, 6LoWPAN and other wireless technologies so that they can support the various data acquisition requirements.

3.2. *Data preprocessing and storage*

3.2.1. Data preprocessing

Data acquired from MIoT has the following characteristics:

- *Heterogeneous data types.* The whole manufacturing chain generates various data types including sensory data, RFID readings, product records, text, logs, audio, video, etc. The data is in the forms of structured, semi-structured and non-structured.
- *Erroneous and noisy data.* The data obtained from industrial environment is often erroneous and noisy mainly due to the following reasons: (a) interference during the process of data collection especially in industrial environment, (b) the failure and malfunction of sensors or machinery, (c) intermittent loss or outage of wireless or wired communications [Siddiqa et al.(2016)]. For example, wireless communications are often susceptible to harsh industrial environmental factors like blockage, shadowing and fading effects. Moreover, data transmission may fail in industrial WSNs due to the depletion of batteries of sensors or machinery.
- *Data redundancy.* Data generated in MIoT often contain excessively redundant information. For instance, it is shown in [Ertek, Chi, and Zhang(2017)] that there are excessive duplicated RFID readings when multiple RFID tags were scanned by several RFID readers at different time slots. The data redundancy often results in data inconsistency.

Data preprocessing approaches on MIoT data include *data cleaning*, *data integration* and *data compression* as shown in Figure 3. In industrial environment, sensory data is usually uncertain and erroneous due to the depletion of battery power of sensors, imprecise measurement of sensors and communication failures. There are several approaches proposed to address these issues. For example, [Zhong et al.(2015)] proposed RFID-Cuboids approach to remove redundant readings and eliminate the missing values. Moreover, an Indoor RFID Multi-variate Hidden Markov Model (IR-MHMM) was proposed to determine uncertain data and remove duplicated RFID readings as shown in [Baba et al.(2017)]. Furthermore, a machine-learning based method was proposed to filter out the invalid RFID readings [Ma, Wang, and Wang(2018)]. In addition, the study of [Bhandari et al.(2017)] proposed an auto-correlation based scheme to remove duplicated time-series temperature data. In [Tasnim, Pissinou, and Iyengar(2017)], a

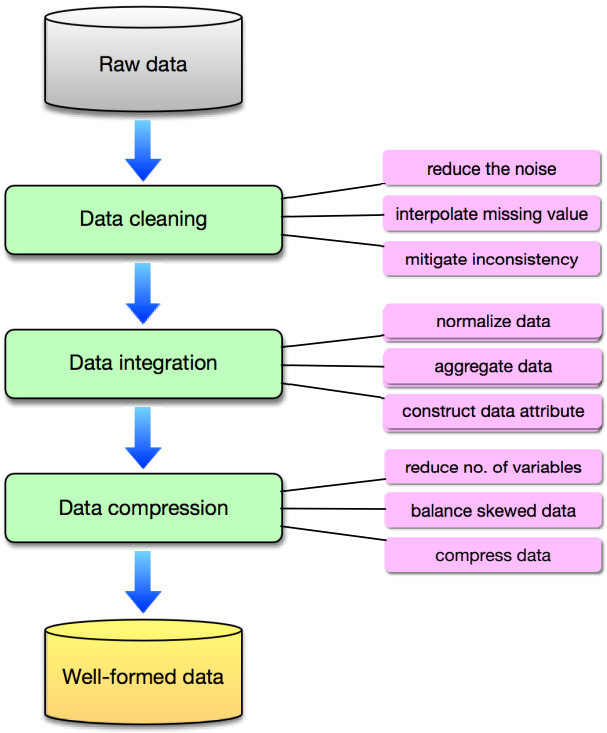


Figure 3. Data preprocessing techniques

novel data cleaning mechanism was proposed to clean erroneous data in environmental sensing applications. Besides duplicated readings, there also exist missing values in MIoT data. In [Zheng et al.(2018)], an interpolation method was proposed to recover the missing values of smart grids data. Moreover, energy-saving is a critical issue in data-cleaning algorithms used in MIoT. In [Deng et al.(2018)], an energy-efficient data-cleaning scheme was proposed.

3.2.2. Data storage

Data storage plays an important role in big data analytics for MIoT. We summarize the solutions of data storage in two aspects: 1) storage infrastructure and 2) data management software.

Storage infrastructure consists of a number of interconnected storage devices. Storage devices typically include: magnetic Harddisk Drive, Solid-State Drives, magnetic taps, USB flash drives, Secure Digital (SD) cards, micro SD cards, Read-Only-Memory (ROM), CD-ROMs, DVD-ROMs, etc. These storage devices can be connected together (via wired or wireless connections) to form the storage infrastructure for MIoT in industrial environment.

Besides storage infrastructure, *data management software* plays an important role in constructing the scalable, effective, reliable storage system to support big data analytics in MIoT. As shown in Figure 1, the data management software consists of three layered components:

- *Distributed file systems.* Google File System (GFS) was proposed and developed by Google [Ghemawat, Gobioff, and Leung(2003)] to support the large data intensive distributed applications such as search engine. Moreover, Hadoop Dis-

tributed File System (HDFS) was proposed by Apache [Shvachko et al.(2010)] as an alternative to GFS. In addition, there are other distributed file systems, such as C# Open Source Managed Operating System (Cosmos) proposed by Microsoft [Chaiken et al.(2008)], XtreemFS [Hupfeld et al.(2008)] and Haystack proposed by Facebook [Beaver et al.(2010)]. Most of them can partially or fully support the storage of large scale data sets. Therefore, most of them can offer the support for large scale data storage of MIoT data.

- *Database management systems (DBMS).* DBMS offers a solution to organize the data in an efficient and effective manner. DBMS software tools can be roughly categorized into two types: traditional relational DBMS (aka SQL databases) and non-relational DBMS (aka Non-SQL databases). SQL databases have been a primary data management approach, especially useful to Material Requirements Planning (MRP), Supply Chain Management (SCM), Enterprise Resource Planning (ERP) in the whole manufacturing chain. Typical SQL databases including commercial databases, such as Oracle, Microsoft SQL server and IBM DB2, and open-source alternatives, such as MySQL, PostgreSQL and SQLite. SQL databases usually store data in tables of records (or rows). This storage method neverthless leads to the poor scalability of databases. For example, when data grows, it is necessary to distribute the load among multiple servers. One of benefits of SQL databases is that most of SQL databases can guarantee ACID (Atomicity, Consistency, Isolation, Durability) properties of database transactions, which is crucial to many commercial applications (e.g., ERP and inventory management). Different from SQL databases, NoSQL databases support various types of data, such as records, text, and binary objects. Compared with traditional relational databases, most of NoSQL databases are usually highly scalable and can support the tremendous amount of data. Therefore, NoSQL databases are promising in managing sensory data, device data, RFID trajectory data in MIoT [Lade, Ghosh, and Srinivasan(2017)].
- *Distributed computing models.* There are a number of distributed computing models proposed for big data analytics. For example, Google MapReduce [Dean and Ghemawat(2008)] is one of the typical programming models used for processing large data sets. Hadoop MapReduce [Apache(2014)] is the open source implementation of Google MapReduce. MapReduce is suffering from the lack of iterations or recursions, which are however required by many data analytics applications, such as data mining, graph analysis and social network analysis. There are some extensions to MapReduce to address this concern, including HaLoop [Bu et al.(2010)], Berkeley Orders of Magnitude (BOOM) Analysis [Alvaro et al.(2010)], Twister [Ekanayake et al.(2010)], iHadoop [Elnikety, Elsayed, and Ramadan(2011)] and iMapReduce [Zhang et al.(2012)]. In addition to MapReduce, there are other alternatives such as Dryad [Isard et al.(2007)], Nephele/PACTs system [Battré et al.(2010)], Spark [Zaharia et al.(2010)], Pregel [Malewicz(2010)], Hive [Thusoo et al.(2010)], GraphLab [Low et al.(2012)].
- *Virtual machines and containers.* Virtual machines (VMs) have been widely used to support cloud computing. Through virtualization, multiple VMs can be emulated on a single computer system. VMs can help to achieve the isolation of multiple virtual operating systems, on top of which multiple applications can be supported. Different from VMs, containers run on top of a single operating system and a single hardware while containers separate the applications as well as the underneath binary and library files. Therefore, containers can achieve the lightweight virtualization, consequently resulting the super fast booting speed,

small size, less resource consumption (compared with VMs). The lightweight features of containers lead to the feasibility to edge computing scenarios (to be illustrated in Section 3.4).

3.3. *Data analytics*

3.3.1. Typical data analytics approaches

Typical data analytics approaches include: 1) *Statistical modeling* schemes, 2) *Data mining* schemes, 3) *Machine learning* schemes and 4) *Data visualization*.

Statistical modeling methods are mainly based on statistical theory. There are three types of statistical methods: (i) descriptive statistics that is used to quantify relationships in data [Trochim, Donnelly, and Arora(2016)]; (ii) inferential statistics that is used to to deduce generalizations from the sample data sets [Bandyopadhyay and Forster(2011)]; (iii) stochastic modeling methods can capture the dynamic features of data traffic, predict user mobility and track objects [Newson and Krumm(2009), Liao et al.(2018)].

Data mining is the process of extracting useful information from massive data sets. There are a wide variety of data mining algorithms that can be used in MIoT such as Apriori algorithm, Frequent Pattern Growth (FP-Growth) algorithm, Density-based spatial clustering of applications with noise (DBSCAN), Generalized Sequential Pattern (GSP), Sequential Pattern Discovery Using Equivalent Class (SPADE) and Prefix-Projected Sequential Pattern Mining (PrefixSpan) [Han, Kamber, and Pei(2012)].

Machine learning explores to construct self-adaptive algorithms that can learn from existing data and perform predictive analysis. As one of typical applications of machine learning, data mining has emphasis on extracting valuable information from data. Typical Machine learning algorithms include support vector machines (SVMs) [Vapnik(1995)], naive Bayes [Wu et al.(2008)], Decision tree learning [Russell and Norvig(2009)], k-Nearest Neighbors (k-NN) [Altman(1992)], hidden Markov model, Bayesian networks [Qiu et al.(2016)], neural networks [Zhang(2000)], Ensemble methods [Zhou(2012)], k-means [Kanungo et al.(2002)], singular value decomposition (SVD), Principal Component Analysis (PCA) [Jolliffe(2002)] and reinforcement learning algorithms such as Q-learning [Russell and Norvig(2009)].

3.3.2. Taxonomy of data analytics approaches in MIoT

We next present an overview of data analytics in MIoT in the aspect of MIoT applications. In particular, data analytics methods in MIoT can be roughly categorized into: 1) Descriptive analytics, 2) Diagnostic analytics, 3) Predictive analytics, 4) Prescriptive analytics. This classification can better represent the data analytics in MIoT applications in different levels of complexity and extracted values. Figure 4 depicts different levels of data analytics methods in MIoT applications. Both descriptive and diagnostic analytics methods are reactive while predictive and prescriptive analytics approaches are proactive. Moreover, prescriptive and predictive analytics approaches are more complicated than descriptive and diagnostic analytics methods though they can bring more values than descriptive and diagnostic analytics. We then present an overview of existing studies in the four levels of data analytics.

(1) Descriptive analytics

Descriptive analytics is an exploratory analysis of historical data to tell what happened. During this stage, most of data mining and statistic methods can be used to

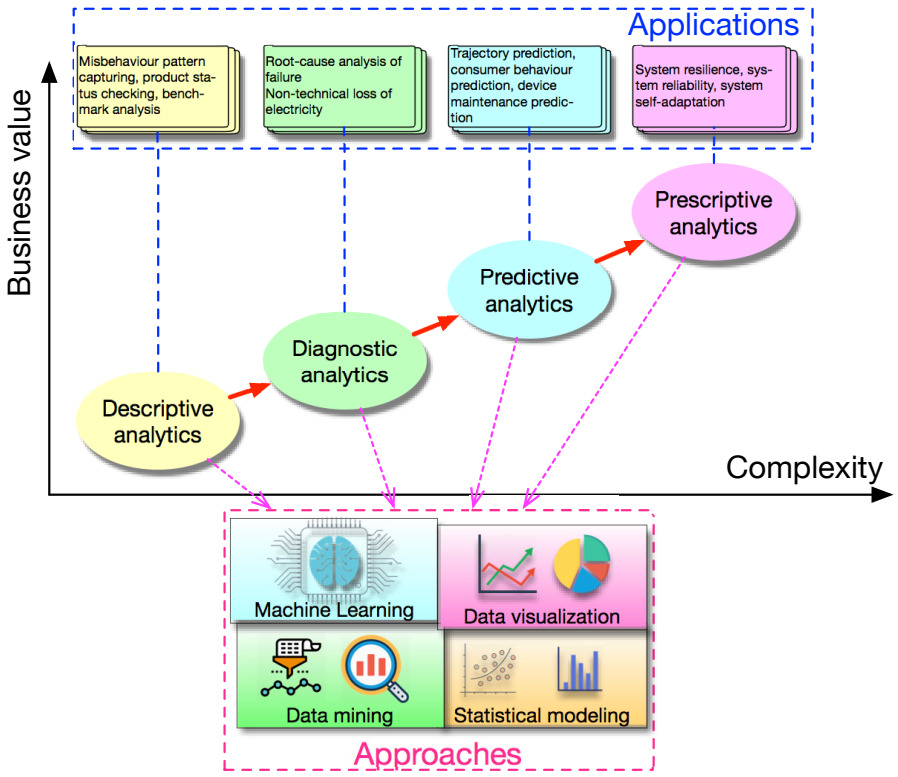


Figure 4. Data analytics

reveal the data characteristics, recognize patterns and identify relationships of data objects. Descriptive analytics can be used in the whole life cycle of manufacturing data. In particular, a real-time monitoring system was proposed in [Zhang et al.(2015)] to track the different manufacturing resources. Zhong et al. [Zhong et al.(2016)] proposed RFID-Cuboid framework to integrate production logistic data with RFID data and offered a system prototype to visualize logistic trajectory data. Moreover, the study of [Zuo, Tao, and Nee(2018)] presented a cloud-based approach to evaluate the energy consumption during product manufacturing process. In addition, air-qualtiy monitoring system based on wireless sensor networks at a logistics shipping base was proposed in [Molka-Danielsen, Engelseth, and Wang(2018)].

(2) Diagnostic analytics

Diagnostic analytics is a deeper look at data to attempt to understand the causes of events and behaviours. The diagnostic analysis of machines and other equipments can help to identify the possible faults and predict the failures to reduce the machine down-times. For example, a method of integrating SVM and artificial neural network (ANN) was presented to detect and diagnose machinery faults of centrifugal pumps [Azadeh et al.(2013)]. The study of [Wang et al.(2016)] proposed fault detection methods for propeller ventilation of vessels based on Kalman filter. Wuest et al. put forth a surpervised maching learning method to monitor product quality in [Wuest, Irgens, and Thoben(2014)]. Compared with supervised machine learning methods, unsupervised learning methods require less feature engineering efforts in obtaining features consequently saving the time and the labor. In [Lei et al.(2016)], a two-stage unsupervised learning method was proposed to conduct diagnostic analysis of machine faults. In addition to fault diagnosis, *anomaly detection* (or outlier detec-

tion) is to identify data objects that do not comply with an expected pattern as given. In [Zheng et al.(2018)], a deep learning based method was proposed to detect electric theft via anomaly detection of electricity consumption data in smart grids.

(3) Predictive analytics

Predictive analytics mainly utilizes historical data to anticipate the trends of data (i.e., what will occur in the future). In [Wu et al.(2017a)], a random forests (RFs) based method was proposed to predict the tool (machine) wear in manufacturing cycle. It is also shown in [Wu et al.(2017a)] that RFs method outperforms ANN and SVMs in terms of prediction accuracy. One of challenges in data analytics of MIoT data is the imbalanced number of negative and postive samples [Lade, Ghosh, and Srinivasan(2017)]. The study of [Kim et al.(2017)] proposed a cost-sensitive decision tree ensemble algorithm to address this issue. Extensive experimental results show that the proposed method outperforms other existing baseline methods. Moreover, in [Ren, Hung, and Tan(2018)], a deep-learning based method was proposed to predict product surface defects. In addition, consumer behaviour prediction plays an important role in manufacturing business stage, e.g., to improve the consumers' purchase decision-makeing predictions. In [Zuo(2016)], a Bayesian network based approach was proposed to predict the customer purchase behaviour. In particular, the analysis is based on massive RFID data, which was collected through RFID tags attached at customers.

(4) Prescriptive analytics

Prescriptive analytics extends the results of descriptive, diagnostic and predictive analytics to make right decisions in order to achieve predicted outcomes (i.e., what should we do to achieve the goal?). The prescriptive methods typically include *simulation*, *decision-making*, *optimization* and *reinforcement learning algorithms*. In particular, in [Gerlach, Hass, and Mandenius(2015)], a conceptual design approach was proposed to simulate the configuration and procedural training in a bio-ethanol plant. The study of [Mourtzis et al.(2016)] presents a novel method for manufacturing-networks design via intelligent decision-making on selecting suppliers to fulfill the requirements of frugal innovation. In [Kluczek(2016)], an analytic hierarchy process (AHP) based method was proposed to evaluate manufacturing sustainability performance. Moreover, in [Drakaki and Tzionas(2017)], a novel method with the integration of Timed Colored Petri Nets (CTPNs) and reinforcement learning (RL) was proposed to solve the problem of manufacturing scheduling.

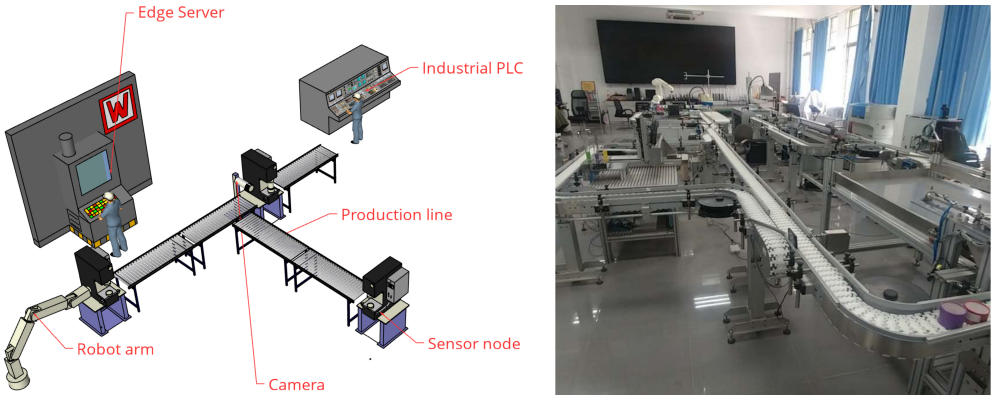
Table 2 summarizes data analytics methods used for MIoT. We categorize them into four types according to different levels in terms of complexity and extracted values. Moreover, we also enumerate representative data analytics methods in each category. In addition, we also list representative application cases in each category.

3.3.3. Data visualization in MIoT

In addition to the aforementioned data analytics, data visualization is also an important tool in MIoT data. Effective data visualization procedure can help to extract and interpret the informative values from complex and high-dimensional MIoT data [Telea(2014)]. Typical data visualization methods include information visualization, exploratory data analysis, statistic plots. The typical quantitative messages that are conveyed by data visualization include: time-series, ranking, frequency distribution, deviation, correlation, part-to-whole, geographic [Post(2003)]. The basic data visualization techniques include: 1) various statistic plots (e.g., bar chart, histogram, pie diagram, scatter plots), 2) word clouds of text data, 3) correlation coefficient matrices/functions, 4) network/graph diagrams of non-structural data, 5) heat map of

Table 2. Classification of data analytics approaches in MIoT

	Questions	Approaches	Applications	References
Descriptive	What happened?	• Association rule mining • Clustering, sequential pattern mining • Querying, statistic reporting • Data visualization	Misbehaviour pattern capturing Product status checking Benchmark analysis	[Zhang et al.(2015)] [Zhong et al.(2016)] [Zuo, Tao, and Nee(2018)] [Molka-Danielsen, Engelseth, and Wang(2018)]
Diagnostic	Why it happened?	• Reasoning • Bayesian analysis	Fault diagnosis Root-cause analysis of failure Anomaly detection	[Azadeh et al.(2013)] [Wang et al.(2016)] [Wuest, Irgens, and Thoben(2014)] [Lei et al.(2016)] [Zheng et al.(2018)]
Predictive	What might happen in the future?	• Classification, regression • Machine learning (supervised /unsupervised) • Deep learning	Trajectory prediction Consumer behaviour prediction Device maintenance prediction	[Wu et al.(2017a)] [Kim et al.(2017)] [Ren, Hung, and Tan(2018)] [Zuo(2016)]
Prescriptive	What should be done?	• Simulation • Optimization • Reinforcement learning (e.g., Q-Learning) • Decision making: e.g., Analytic Hierarchy Process (AHP), The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)	System resilience System reliability System optimization	[Gerlach, Hass, and Mandenius(2015)] [Mourtzis et al.(2016)] [Kluczek(2016)] [Drakaki and Tzionas(2017)]



(a) System Prototype (b) Realistic deployment of system prototype

Figure 5. Case study for distributed computing models for MIoT

geographic data.

Typical data visualization toolboxes include Matlab plot (`https://it.mathworks.com/help/matlab/ref/plot.html`), gnuplot (`http://www.gnuplot.info/`), Python's Seaborn (`https://seaborn.pydata.org/`), Pandas plot (`https://pandas.pydata.org/`), Matplotlib (`https://matplotlib.org/`). Moreover, web-based visualization tools have also been wide used. Representative web-based data visualization tools include Tableau (`https://www.tableau.com/`), Plotly (`https://plot.ly/`), Sisense (`https://www.sisense.com/`), D3.js (`https://d3js.org/`).

3.4. *Case studies*

To demonstrate the feasibility of distributed computing models in MIoT, we developed a system prototype. Figure 5(a) shows that the system framework consists of a production line, industrial devices and computing units. In particular, the production line consists of various manufacturing devices, instruments, sensors, actuators and robot arms, all of which are connected through wired or wireless links consequently forming the MIoT. In addition to the production line and industrial devices, there are a number computing units supporting diverse data processing tasks. For example, edge computing servers with equipped with embedded computers are deployed in the proximity to MIoT. Moreover, the computing-intensive tasks may be uploaded to the remote cloud servers while the latency-sensitive tasks may be processed at edge servers.

In the computing perspective, we develop a distributed computing platform with the orchestration of remote cloud computing and local edge computing. In particular, we deploy Xen hypervisor at remote cloud servers and Docker container at edge servers. On top of virtual machines, we further utilize Hadoop distributed computing platforms to support big data processing tasks. In order to coordinate the edge and cloud computing tasks, we design and implement a hybrid edge/cloud computing framework (details can be referred to the work [Li et al.(2019)]).

Figure 5(b) gives the realistic prototype of a printed circuit board (PCB) production line based on our proposed system framework. This production line consists of conveyor belts, product feeding machines, robot arms, sensors and cameras. We choose industrial WLANs as the wired connections and 6LoWPAN as the wireless connections. In addition, we adopt 4 edge servers, each of which has the identical configurations:

a single-board computer with a quad-core Broadcom BCM2837 CPU, 1GB memory and 64GB SSD storage. Furthermore, there is a remote cloud server (i.e., IBM X3650 M3) with 2 Intel Xeon Processors, 24 GB memory and 1TB SSD storage.

We then evaluate the performance of the proposed hybrid edge/cloud computing framework on top of the prototype. In particular, we consider a pure cloud computing framework and a pure edge computing framework as baseline models. Moreover, image recognition tasks with varied image size were chosen to be executed at edge and cloud servers. We further adopt OpenCV frameworks on both edge and cloud servers to support the image recognition tasks.

Table 3. Performance evaluation

	10 MB	12 MB	14 MB	16 MB	18 MB	20 MB
Cloud Computing Only (second)	1.20	1.48	1.67	1.82	2.08	2.45
Edge Computing Only (second)	**0.61**	**0.86**	**0.97**	1.15	1.26	1.43
Hybrid Cloud and Edge (second)	0.75	0.93	0.98	**0.86**	**0.97**	**0.96**

Table 3 shows the latency values of three computing frameworks versus varied image sizes. In particular, the latency is calculated via averaging results with 100 images, each with the same image size (e.g., 10 MB). It is shown in Table 3 that the average latency is increased with the increased image size; this effect may owe to the increased computational complexity of image recognition algorithms with the increased image size. We also observe from Table 3 that the proposed hybrid cloud and edge scheme outperforms pure cloud computing scheme and pure edge computing scheme with larger image size (e.g., 16 MB, 18 MB and 20 MB). It can be explained as follows: 1) pure cloud computing has the strength in processing large images while suffering from the long end-to-end latency; 2) pure edge computing scheme can complete the computing tasks with smaller image size (e.g., 12 MB) and achieve the short end-to-end latency due to the deployment proximity; 3) hybrid edge/cloud computing scheme can not only exploit the strength of cloud computing to process the complicated tasks but also harness the benefit of edge computing in short latency, consequently obtaining the better performance in the cases with larger image size.

4. Future research directions

In this section, we discuss open issues as well as future directions in big data analytics for MIoT. Figure 6 summarizes the future directions in big data analytics in MIoT.

4.1. *Security and Privacy Concerns*

Privacy and security are becoming an arising challenge of big data analytics for MIoT. Privacy concerns the proper utilization of the data with the preservation of enterprise private information, whereas security is to ensure data confidentiality, integrity and availability [Wang et al.(2018b)]. We next summarize the research issues related to privacy and security in big data analytics for MIoT.

- *Security assurance in data acquisition.* The proliferation of wireless connections in manufacturing industry results in the challenges in security assurance during data acquisition because of the openness of wireless medium susceptible to ma-

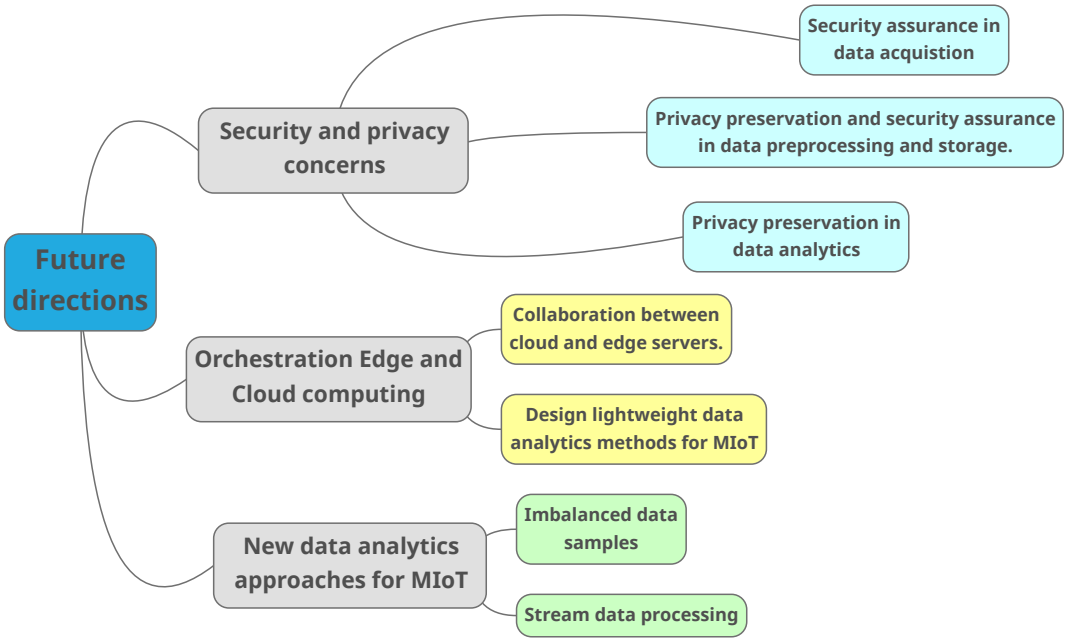


Figure 6. Future directions in big data analytics of MIoT

licious attacks like passive eavesdropping attacks [Li et al.(2018)]. The typical countermeasure is to apply encryption schemes in wireless networks [Hennebert and Santos(2014)]. However, it may not be feasible to apply cryptography-based techniques in all IoT networks due to the following constraints: the inferior computational capability and the limited battery power of some smart objects like RFID and sensors. Therefore, new protection schemes without strong computational complexity and high energy consumption shall be developed for MIoT in the future.

- *Privacy preservation and security assurance in data preprocessing and storage.* After data acquisition, MIoT data will be preprocessed and stored locally (at servers of factories or other departments) or remotely (at remote cloud servers) [Wang, Gao, and Fan(2015)]. However, the distribution of MIoT data throughout the enterprise consisting of multiple manufacturing sites across different regions often results in the vulnerability to various malicious attacks from insiders and outsiders of the enterprise. It is challenging to offer a solution against malicious attacks. There are several possible directions in solving this issue: 1) Proper key management [Esposito et al.(2016)] including proper key distribution and key validation period, 2) authentication mechanism including accessing control of files and data records, 3) traceability of data accessing allowing any data accessing or modification to be identifiable so that the malicious behaviours can be avoided or revoked.
- *Privacy preservation in data analytics.* In order to protect data privacy, the data is often encrypted and stored at a server (or at a cloud). Before data analytics, the data needs to be decrypted. However, the decryption process is often time-consuming consequently resulting in the inefficiency of data analytics in MIoT. How to design a privacy-preservation scheme of balancing the efficiency and

privacy becomes a challenge [Wang et al.(2018b), Babar et al.(2019)].

4.2. *Edge Computing for big data analytics in MIoT*

The integration of cloud computing with manufacturing brings the opportunities in saving the capital investments of information and communication technologies (ICT), providing flexibility of ICT resources to small and medium enterprises [Wang, Gao, and Fan(2015), Esposito et al.(2016)]. However, there are also limitations with cloud computing such as high latency, performance bottleneck, single-point-to-failure and privacy leakage [Liu et al.(2017)]. Recently, mobile edge computing (or fog computing) has become a new complement to cloud computing by offloading both *computational and storage tasks* from remote cloud servers to local edge servers [Tran et al.(2017), Wu et al.(2017b), Wang et al.(2017)]. In this manner, the computing-intensive and delay-tolerant tasks will be executed at remote cloud servers while the delay-critical and computing less-intensive tasks will be offloaded to edge servers. As a result, the real-time tasks like sensing, monitoring and controlling can be enabled in the proximity to factories and enterprises. The case study in Section 3.4 also demonstrates the effectiveness of hybrid edge and cloud computing in MIoT.

However, there are many challenges in edge computing for big data analytics in MIoT.

- *Collaboration between cloud and edge servers.* There are diversity of computing resources in manufacturing networks. For example, remote cloud servers usually have superior computing capability than local edge servers while there is a longer delay to upload the tasks to the remote cloud servers than to upload the tasks to the local edge servers Therefore, it is necessary to determine how to allocate the computational tasks at cloud servers or at edge servers. For example, the computing intensive and delay-tolerant tasks should be uploaded to remote cloud servers while the computing less-intensive and delay-critical tasks can be executed locally at edge servers. In this sense, edge servers can be deployed within factories and remote clouds can be deployed outside factories (even if they can be provided by third parties). To the best of our knowledge, there are few studies on investigating collaboration between cloud and edge servers, especially in the whole manufacturing network. In the future, research efforts should be done in allocating and coordinating various computing resources distributed in cloud and edge servers in manufacturing.
- *Design lightweight data analytics methods for MIoT.* Many data analytics tasks that are delay-critical should be executed locally at edge servers (or at manufacturing devices). However, due to the resource limitation of edge severs, the conventional data analytics methods might be too complicated to be executed at edge servers. Therefore, the models of the data analytics methods need to be trained at remote cloud servers first and be transferred at local edge servers. However, it can result in huge communication cost to transmit this model from the remote cloud servers to the edge servers. For example, the study of [Lin et al.(2018)] shows that AlexNet (i.e., a typical deep learning method) has the model size of 240MB, which is so large that it can cause extra delay from the cloud server to the edge server. Therefore, it is necessary to design lightweight data analytics schemes which can be deployed locally at edge servers approximate to users [Leng et al.(2018a)].

4.3. *New data analytics methods for MIoT data*

Although a lot of efforts have been done in developing data analytics methods for MIoT data, there are still many open research issues in this area.

- *Imbalanced data samples.* Different from data analytics in traditional fields (e.g., commercial database systems), manufacturing data has the imbalanced number of data samples between positive and negative samples. For example, it is shown in [Lade, Ghosh, and Srinivasan(2017)] that the ratio of positive samples to negative samples (vice versa) can be 99,000,000 to 1. It is challenging to apply conventional data analytics methods to analyse the imbalanced dataset. Therefore, new data analytics methods should be developed to solve this issue. To the best of our knowledge, there are few studies [Kim et al.(2017)] proposed to address this issue.
- *Stream data processing.* In MIoT, there is a tremendous volume of real-time data generated (e.g., sensory data from industrial wireless sensor networks) [Wang et al.(2018a)]. It is impossible to store and process the entire data in the memory of computers. Consequently, the conventional methods requiring saving the whole data sets in memory cannot work in this scenario. It is challenging to analyse the massive data-stream of MIoT. It is worthwhile to investigate new data analytics approaches to process the data-stream of MIoT.

5. Conclusion

This paper presents an in-depth survey on big data analytics in manufacturing Internet of Things (MIoT). This paper first presents a life cycle of big data analytics in MIoT and discusses the necessities as well as challenges of big data analytics in MIoT. Then, the enabling technologies of big data analytics in MIoT are summarized according to three phases in the life cycle of big data analytics: data acquisition, data preprocessing and storage, and data analytics. Moreover, this paper also outlines the future directions and discusses the open research issues. We believe big data analytics will play an important role in promoting manufacturing industry to evolve into smart manufacturing in the foreseeable future.

Disclosure statement

The authors declare that they have no potential conflict of interest.

References

[Altman(1992)] Altman, Naomi S. 1992. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician* 46 (3): 175–185.

[Alvaro et al.(2010)] Alvaro, Peter, Tyson Condie, Neil Conway, Khaled Elmeleegy, Joseph M. Hellerstein, and Russell Sears. 2010. "Boom Analytics: Exploring Data-centric, Declarative Programming for the Cloud." In *Proceedings of the 5th European Conference on Computer Systems (EuroSys)*, .

[Apache(2014)] Apache. 2014. "Hadoop MapReduce." `https://hadoop.apache.org/`.

[Azadeh et al.(2013)] Azadeh, A., M. Saberi, A. Kazem, V. Ebrahimipour, A. Nourmohammadzadeh, and Z. Saberi. 2013. "A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization." *Applied Soft Computing* 13 (3): 1478 – 1485.

[Azoidou et al.(2017)] Azoidou, E., Z. Pang, Y. Liu, D. Lan, G. Bag, and S. Gong. 2017. "Battery Lifetime Modeling and Validation of Wireless Building Automation Devices in Thread." *IEEE Transactions on Industrial Informatics* .

[Baba et al.(2017)] Baba, Asif Iqbal, Hua Lu, Torben Bach Pedersen, and Manfred Jaeger. 2017. "Cleansing Indoor RFID Tracking Data." *SIGSPATIAL Special* 9 (1): 11–18.

[Babar et al.(2019)] Babar, Muhammad, Fahim Arif, Mian Ahmad Jan, Zhiyuan Tan, and Fazlullah Khan. 2019. "Urban data management system: Towards Big Data analytics for Internet of Things based smart urban environment using customized Hadoop." *Future Generation Computer Systems* 96: 398 – 409. `http://www.sciencedirect.com/science/article/pii/S0167739X18321095`.

[Bandyopadhyay and Forster(2011)] Bandyopadhyay, Prasanta S., and Malcolm R. Forster. 2011. *Philosophy of Statistics*. Elsevier.

[Battré et al.(2010)] Battré, Dominic, Stephan Ewen, Fabian Hueske, Odej Kao, Volker Markl, and Daniel Warneke. 2010. "Nephele/PACTs: A Programming Model and Execution Framework for Web-scale Analytical Processing (SoCC)." In *Proceedings of the 1st ACM Symposium on Cloud Computing*, .

[Beaver et al.(2010)] Beaver, Doug, Sanjeev Kumar, Harry C. Li, Jason Sobel, and Peter Vajgel. 2010. "Finding a Needle in Haystack: Facebook's Photo Storage." In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, .

[Bhandari et al.(2017)] Bhandari, Siddhartha, Neil Bergmann, Raja Jurdak, and Branislav Kusy. 2017. "Time Series Data Analysis of Wireless Sensor Network Measurements of Temperature." *Sensors* 17 (6).

[Bu et al.(2010)] Bu, Yingyi, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. 2010. "HaLoop: Efficient Iterative Data Processing on Large Clusters." *Proc. VLDB Endow.* 3 (1-2).

[Casado and Younas(2015)] Casado, Rubén, and Muhammad Younas. 2015. "Emerging Trends and Technologies in Big Data Processing." *Concurr. Comput. : Pract. Exper.* 27 (8): 2078–2091.

[Chaiken et al.(2008)] Chaiken, Ronnie, Bob Jenkins, Per-Å ke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, and Jingren Zhou. 2008. "SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets." *Proc. VLDB Endow.* 1 (2): 1265–1276.

[Chi et al.(2014)] Chi, Q., H. Yan, C. Zhang, Z. Pang, and L. D. Xu. 2014. "A Reconfigurable Smart Sensor Interface for Industrial WSN in IoT Environment." *IEEE Transactions on Industrial Informatics* 10 (2): 1417–1425.

[Dean and Ghemawat(2008)] Dean, Jeffrey, and Sanjay Ghemawat. 2008. "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM* 51 (1): 107–113.

[Deng et al.(2018)] Deng, Changyi, Ruifeng Guo, Chao Liu, Ray Y. Zhong, and Xun Xu. 2018. "Data cleansing for energy-saving: a case of Cyber-Physical Machine Tools health monitoring system." *International Journal of Production Research* 56 (1-2): 1000–1015.

[Drakaki and Tzionas(2017)] Drakaki, Maria, and Panagiotis Tzionas. 2017. "Man-

ufacturing Scheduling Using Colored Petri Nets and Reinforcement Learning." *Applied Sciences* 7 (2).

[Ekanayake et al.(2010)] Ekanayake, Jaliya, Hui Li, Bingjing Zhang, Thilina Gunarathne, Seung-Hee Bae, Judy Qiu, and Geoffrey Fox. 2010. "Twister: A Runtime for Iterative MapReduce." In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC)*, .

[Elnikety, Elsayed, and Ramadan(2011)] Elnikety, E., T. Elsayed, and H. E. Ramadan. 2011. "iHadoop: Asynchronous Iterations for MapReduce." In *IEEE CloudCom*, .

[Ertek, Chi, and Zhang(2017)] Ertek, G., X. Chi, and A. N. Zhang. 2017. "A Framework for Mining RFID Data From Schedule-Based Systems." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47 (11): 2967–2984.

[Esposito et al.(2016)] Esposito, C., A. Castiglione, B. Martini, and K. K. R. Choo. 2016. "Cloud Manufacturing: Security, Privacy, and Forensic Concerns." *IEEE Cloud Computing* 3 (4): 16–22.

[Gerlach, Hass, and Mandenius(2015)] Gerlach, Inga, Volker C. Hass, and Carl-Fredrik Mandenius. 2015. "Conceptual Design of an Operator Training Simulator for a Bio-Ethanol Plant." *Processes* 3 (3): 664–683.

[Ghemawat, Gobioff, and Leung(2003)] Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. 2003. "The Google File System." In *Proceedings of ACM SOSP*, .

[Guerra et al.(2011)] Guerra, Jorge, Himabindu Pucha, Joseph Glider, Wendy Belluomini, and Raju Rangaswami. 2011. "Cost Effective Storage Using Extent Based Dynamic Tiering." In *Proceedings of the 9th USENIX Conference on File and Stroage Technologies (FAST)*, .

[Han, Kamber, and Pei(2012)] Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques*. Third edition ed. Boston, USA: Morgan Kaufmann.

[Hennebert and Santos(2014)] Hennebert, C., and J. D. Santos. 2014. "Security Protocols and Privacy Issues into 6LoWPAN Stack: A Synthesis." *IEEE Internet of Things Journal* 1 (5): 384–398.

[Hu et al.(2014)] Hu, H., Y. Wen, T. S. Chua, and X. Li. 2014. "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial." *IEEE Access* 2: 652–687.

[Hupfeld et al.(2008)] Hupfeld, Felix, Toni Cortes, Björn Kolbeck, Jan Stender, Erich Focht, Matthias Hess, Jesus Malo, Jonathan Marti, and Eugenio Cesario. 2008. "The XtreemFS Architecture&Mdash;a Case for Object-based File Systems in Grids." *Concurrency and Computation: Practice and Experience* 20 (17): 2049–2060.

[Isard et al.(2007)] Isard, Michael, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly. 2007. "Dryad: Distributed Data-parallel Programs from Sequential Building Blocks." *SIGOPS Oper. Syst. Rev.* 41 (3): 59–72.

[Jolliffe(2002)] Jolliffe, I.T. 2002. *Principal Component Analysis*. Springer-Verlag.

[Kanungo et al.(2002)] Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. "An Efficient k-Means Clustering Algorithm: Analysis and Implementation." *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (7): 881–892.

[Kim et al.(2017)] Kim, Aekyung, Kyuhyup Oh, Jae-Yoon Jung, and Bohyun Kim. 2017. "Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles." *International Journal of Computer Integrated*

Manufacturing 1–17.

[Kluczek(2016)] Kluczek, Aldona. 2016. "Application of multi-criteria approach for sustainability assessment of manufacturing processes." *Management and Production Engineering Review* 7 (3): 62–78.

[Kusiak(2017)] Kusiak, A. 2017. "Smart manufacturing must embrace big data." *Nature* 544 (7648): 23.

[Kusiak(2018)] Kusiak, Andrew. 2018. "Smart manufacturing." *International Journal of Production Research* 56 (1-2): 508–517.

[Lade, Ghosh, and Srinivasan(2017)] Lade, Prasanth, Rumi Ghosh, and Soundar Srinivasan. 2017. "Manufacturing Analytics and Industrial Internet of Things." *IEEE Intelligent Systems* 32 (3): 74–79.

[Lei et al.(2016)] Lei, Y., F. Jia, J. Lin, S. Xing, and S. X. Ding. 2016. "An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data." *IEEE Transactions on Industrial Electronics* 63 (5): 3137–3147.

[Leng et al.(2018a)] Leng, Cong, Hao Li, Shenghuo Zhu, and Rong Jin. 2018a. "Extremely Low Bit Neural Network: Squeeze the Last Bit Out with ADMM." In *AAAI*, .

[Leng et al.(2018b)] Leng, Kaijun, Linbo Jin, Wen Shi, and Inneke Van Nieuwenhuyse. 2018b. "Research on agricultural products supply chain inspection system based on internet of things." *Cluster Computing* .

[Li et al.(2019)] Li, X., J. Wan, H. Dai, M. Imran, M. Xia, and A. Celesti. 2019. "A Hybrid Computing Solution and Resource Scheduling Strategy for Edge Computing in Smart Manufacturing." *IEEE Transactions on Industrial Informatics* (early access): 1–9.

[Li et al.(2018)] Li, Xuran, Qiu Wang, Hong-Ning Dai, and Hao Wang. 2018. "A Novel Friendly Jamming Scheme in Industrial Crowdsensing Networks against Eavesdropping Attack." *Sensors* 18 (6).

[Liao et al.(2018)] Liao, Yongxin, Hervé Panetto, Paulo C. Stadzisz, and Jean M. Simão. 2018. "A notification-oriented solution for data-intensive enterprise information systems – A cloud manufacturing case." *Enterprise Information Systems* 12 (8-9): 942–959.

[Lin et al.(2018)] Lin, Yujun, Song Han, Huizi Mao, Yu Wang, and William J. Dally. 2018. "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training." In *International Conference on Learning Representations (ICLR)*, .

[Liu et al.(2017)] Liu, H., F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang. 2017. "Mobile Edge Cloud System: Architectures, Challenges, and Approaches." *IEEE Systems Journal* PP (99): 1–14.

[Low et al.(2012)] Low, Yucheng, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M. Hellerstein. 2012. "Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud." *Proc. VLDB Endow.* 5 (8): 716–727.

[Ma, Wang, and Wang(2018)] Ma, Haishu, Yi Wang, and Kesheng Wang. 2018. "Automatic detection of false positive RFID readings using machine learning algorithms." *Expert Systems with Applications* 91: 442 – 451.

[Malewicz(2010)] Malewicz, Grzegorz et al. 2010. "Pregel: A System for Large-scale Graph Processing." In *Proceedings of ACM SIGMOD*, .

[Mekki et al.(2018)] Mekki, Kais, Eddy Bajic, Frederic Chaxel, and Fernand Meyer. 2018. "A comparative study of LPWAN technologies for large-scale IoT deployment." *ICT Express* .

[Molka-Danielsen, Engelseth, and Wang(2018)] Molka-Danielsen, Judith, Per Engelseth, and Hao Wang. 2018. "Large scale integration of wireless sensor network technologies for air quality monitoring at a logistics shipping base." *Journal of Industrial Information Integration* 10: 20 – 28. http://www.sciencedirect.com/science/article/pii/S2452414X17301024.

[Mourtzis et al.(2016)] Mourtzis, D., E. Vlachou, N. Boli, L. Gravias, and C. Giannoulis. 2016. "Manufacturing Networks Design through Smart Decision Making towards Frugal Innovation." *Procedia CIRP* 50: 354 – 359. 26th CIRP Design Conference, http://www.sciencedirect.com/science/article/pii/S2212827116304061.

[Newson and Krumm(2009)] Newson, Paul, and John Krumm. 2009. "Hidden Markov Map Matching Through Noise and Sparseness." In *Proceedings of ACM SIGSPATIAL*, .

[Petersen and Carlsen(2011)] Petersen, S., and S. Carlsen. 2011. "WirelessHART Versus ISA100.11a: The Format War Hits the Factory Floor." *IEEE Industrial Electronics Magazine* 5 (4): 23–34.

[Post(2003)] Post, Nielson Gregory Bonneau Georges-Pierre, Frits H. 2003. *Data Visualization - the State of the Art*. Springer-Verlag.

[Qiu et al.(2016)] Qiu, Junfei, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. "A survey of machine learning for big data processing." *EURASIP Journal on Advances in Signal Processing* 2016 (1): 1–16.

[Ren, Hung, and Tan(2018)] Ren, R., T. Hung, and K. C. Tan. 2018. "A Generic Deep-Learning-Based Approach for Automated Surface Inspection." *IEEE Transactions on Cybernetics* 48 (3): 929–940.

[Russell and Norvig(2009)] Russell, Stuart, and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach (3rd Edition)*. 3rd ed. Prentice Hall.

[Shvachko et al.(2010)] Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. "The Hadoop Distributed File System." In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, .

[Siddiqa et al.(2016)] Siddiqa, Aisha, Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Mohsen Marjani, Shahabuddin Shamshirband, Abdullah Gani, and Fariza Nasaruddin. 2016. "A survey of big data management: Taxonomy and state-of-the-art." *Journal of Network and Computer Applications* 71: 151 – 166. http://www.sciencedirect.com/science/article/pii/S1084804516300583.

[Tao and Qi(2019)] Tao, F., and Q. Qi. 2019. "New IT Driven Service-Oriented Smart Manufacturing: Framework and Characteristics." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49 (1): 81–91.

[Tao et al.(2018)] Tao, Fei, Qinglin Qi, Ang Liu, and Andrew Kusiak. 2018. "Data-driven smart manufacturing." *Journal of Manufacturing Systems* .

[Tasnim, Pissinou, and Iyengar(2017)] Tasnim, S., N. Pissinou, and S. S. Iyengar. 2017. "A novel cleaning approach of environmental sensing data streams." In *2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 632–633.

[Telea(2014)] Telea, Alexandru C. 2014. *Data visualization: principles and practice*. CRC Press.

[Thusoo et al.(2010)] Thusoo, A., J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy. 2010. "Hive - a petabyte scale data warehouse using Hadoop." In *IEEE 26th International Conference on Data Engineering (ICDE)*, .

[Tran et al.(2017)] Tran, T. X., A. Hajisami, P. Pandey, and D. Pompili. 2017. "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges." *IEEE Communications Magazine* 55 (4): 54–61.

[Trochim, Donnelly, and Arora(2016)] Trochim, William M.K., Jim Donnelly, and Kanika Arora. 2016. *Research Methods The Essential Knowledge Base*. 2nd ed. Cengage Learning.

[Vapnik(1995)] Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

[Wang et al.(2016)] Wang, H., S. Fossen, F. Han, I. A. Hameed, and G. Li. 2016. "Towards data-driven identification and analysis of propeller ventilation." In *OCEANS*, 1–6.

[Wang et al.(2018a)] Wang, Jinjiang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. 2018a. "Deep learning for smart manufacturing: Methods and applications." *Journal of Manufacturing Systems* .

[Wang et al.(2018b)] Wang, Ning, Xiaokui Xiao, Yin Yang, Ta Duy Hoang, Hyejin Shin, Junbum Shin, and Ge Yu. 2018b. "PrivTrie: Effective Frequent Term Discovery under Local Differential Privacy." In *IEEE International Conference on Data Engineering (ICDE)*, .

[Wang, Gao, and Fan(2015)] Wang, Peng, Robert X Gao, and Zhaoyan Fan. 2015. "Cloud computing for cloud manufacturing: benefits and limitations." *Journal of Manufacturing Science and Engineering* 137 (4): 1–9.

[Wang et al.(2018a)] Wang, X., W. Wang, L. T. Yang, S. Liao, D. Yin, and M. J. Deen. 2018a. "A Distributed HOSVD Method With Its Incremental Computation for Big Data in Cyber-Physical-Social Systems." *IEEE Transactions on Computational Social Systems* 5 (2): 481–492.

[Wang et al.(2018b)] Wang, X., L. T. Yang, H. Liu, and M. J. Deen. 2018b. "A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives." *IEEE Transactions on Big Data* 4 (3): 325–340.

[Wang et al.(2017)] Wang, X., L. T. Yang, X. Xie, J. Jin, and M. J. Deen. 2017. "A Cloud-Edge Computing Framework for Cyber-Physical-Social Services." *IEEE Communications Magazine* 55 (11): 80–85.

[Wu et al.(2017a)] Wu, Dazhong, Connor Jennings, Janis Terpenny, Robert X Gao, and Soundar Kumara. 2017a. "A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests." *Journal of Manufacturing Science and Engineering* 139 (7): 071018.

[Wu et al.(2017b)] Wu, Dazhong, Shaopeng Liu, Li Zhang, Janis Terpenny, Robert X. Gao, Thomas Kurfess, and Judith A. Guzzo. 2017b. "A fog computing-based framework for process monitoring and prognosis in cyber-manufacturing." *Journal of Manufacturing Systems* 43: 25 – 34. `http://www.sciencedirect.com/science/article/pii/S0278612517300237`.

[Wu et al.(2008)] Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2008. "Top 10 Algorithms in Data Mining." *Knowl. Inf. Syst.* 14 (1): 1–37.

[Wuest, Irgens, and Thoben(2014)] Wuest, Thorsten, Christopher Irgens, and Klaus-Dieter Thoben. 2014. "An approach to monitoring quality in manufacturing using supervised machine learning on product state data." *Journal of Intelligent Manufacturing* 25 (5): 1167–1180.

[Xu et al.(2017)] Xu, J., J. Yao, L. Wang, Z. Ming, K. Wu, and L. Chen. 2017. "Narrowband Internet of Things: Evolutions, Technologies and Open Issues." *IEEE Internet of Things Journal* PP (99): 1–13.

[Zaharia et al.(2010)] Zaharia, Matei, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. "Spark: Cluster Computing with Working Sets." In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud)*, .

[Zhang(2000)] Zhang, G. P. 2000. "Neural networks for classification: a survey." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30 (4): 451–462.

[Zhang et al.(2012)] Zhang, Yanfeng, Qixin Gao, Lixin Gao, and Cuirong Wang. 2012. "iMapReduce: A Distributed Computing Framework for Iterative Computation." *Journal of Grid Computing* 10 (1): 47–68.

[Zhang et al.(2015)] Zhang, Yingfeng, Geng Zhang, Junqiang Wang, Shudong Sun, Shubin Si, and Teng Yang. 2015. "Real-time information capturing and integration framework of the internet of manufacturing things." *International Journal of Computer Integrated Manufacturing* 28 (8): 811–822.

[Zheng et al.(2018)] Zheng, Z., Y. Yang, X. Niu, H. N. Dai, and Y. Zhou. 2018. "Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids." *IEEE Transactions on Industrial Informatics* 14 (4): 1606–1615.

[Zhong et al.(2015)] Zhong, Ray Y., George Q. Huang, Shulin Lan, Q.Y. Dai, Xu Chen, and T. Zhang. 2015. "A big data approach for logistics trajectory discovery from RFID-enabled production data." *International Journal of Production Economics* 165: 260 – 272. `http://www.sciencedirect.com/science/article/pii/S0925527315000481`.

[Zhong et al.(2016)] Zhong, Ray Y., Shulin Lan, Chen Xu, Qingyun Dai, and George Q. Huang. 2016. "Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing." *The International Journal of Advanced Manufacturing Technology* 84 (1): 5–16. `https://doi.org/10.1007/s00170-015-7702-1`.

[Zhong et al.(2017)] Zhong, Ray Y., Chen Xu, Chao Chen, and George Q. Huang. 2017. "Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors." *International Journal of Production Research* 55 (9): 2610–2621.

[Zhou(2012)] Zhou, Zhi-Hua. 2012. *Ensemble Methods: Foundations and Algorithms.* 1st ed. Chapman & Hall/CRC.

[Zuo(2016)] Zuo, Yi. 2016. "Prediction of Consumer Purchase Behaviour Using Bayesian Network: An Operational Improvement and New Results Based on RFID Data." *Int. J. Knowl. Eng. Soft Data Paradigm.* 5 (2): 85–105.

[Zuo, Tao, and Nee(2018)] Zuo, Ying, Fei Tao, and AYC Nee. 2018. "An Internet of things and cloud-based approach for energy consumption evaluation and analysis for a product." *International Journal of Computer Integrated Manufacturing* 31 (4-5): 337–348.